IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant: Yip

Serial No.: 09/663,968
Confirmation No.: 4499
Filed: September 19, 2000

For: *METHOD AND DEVICE FOR IDENTIFYING A BIOLOGICAL SAMPLE*

Art Unit: 1743

Examiner: Unassigned

ATTACHMENT TO THE PRELIMINARY AMENDMENT MARKED UP PARAGRAPHS AND CLAIMS (37 CFR §1.121)

IN THE SPECIFICATION:

Please amend the specification as follows:

Please amend the paragraph beginning on page 1, lines 1-3, as follows:

[Priority is claimed]This application is related to U.S. patent application serial number 09/285,481, filed April 2, 1999, and entitled "Automated Process Line", which is referred to and incorporated herein in its entirety by this reference.

Please amend the paragraphs beginning on page 2, line 4, through page 3, line 12, as follows:

In a particularly exciting area of genomics, the identification and classification of minute variations in human DNA has been linked with fundamental treatment or medical advice for a specific individual. For example, the variations are a strong indication of predisposition for a particular disease, drug tolerance, and drug efficiency. The most promising of these minute variations are commonly referred to as Single Nucleotide Polymorphisms (SNPs), which relate to a single base-pair change between a first subject and a second subject. By accurately and fully identifying such SNPs, a health care provider would have a powerful indication of a person's likelihood of succumbing to a particular disease, which drugs will be most effective for that person, and what drug treatment plan will be most beneficial. Armed with such knowledge, the health care provider can assist a person in lowering other risk factors for high-susceptibility diseases. Further, the health care provider can confidently select appropriate drug therapies, a process which is now an iterative, hit or miss process where different drugs and treatment schedules are tried until an effective one is found. Not only is this a waste of limited medical resources, but the time lost in finding an effective therapy can have serious medical consequences for the patient.

In order to fully benefit from the use of SNP data, vast quantities of DNA data must be collected, compared, and analyzed. For example, collecting and identifying the SNP profile for a single human subject requires the collection, identification, and classification of thousands, even tens of thousands of DNA samples. Further, the analysis of the resulting DNA data must be carried out with precision. In making a genetic call, where a composition of a biological sample is identified, any error in the call may result in detrimentally affecting the medical advice or treatment given to a [given] patient.

Conventional, known systems and processes for collecting and analyzing DNA data are inadequate to timely and efficiently implement a widespread medical program benefiting from SNP information. For example, many known DNA analysis techniques require the use of an operator or technician to monitor and review the DNA data. An operator, even with sufficient training and substantial experience, is still likely to occasionally make a classification error. For example, the operator may incorrectly identify a base-pair, leading to that patient receiving faulty SNP profile. Alternatively, the operator may view the data and decide that the data [does]do not clearly identify any particular base pair. Although such a "no call" may be warranted, it is likely that the operator will make "no-call" decisions when the data actually [supports]support a valid call. In such a manner, the opportunity to more fully profile the patient is lost.

Please amend the paragraphs on page 3, line 22, through page 4, line 13, as follows:

Briefly, the method and system for identifying a biological sample generates a data set indicative of the composition of the biological sample. In a particular example, the data set is DNA spectrometry data received from a mass spectrometer. The data set is denoised, and a baseline is deleted. Since possible compositions of the biological sample may be known, expected peak areas may be determined. Using the expected peak areas, a residual baseline is generated to further correct the data set. Probable peaks are then identifiable in the corrected data set, which are used to identify the composition of the biological sample. In a disclosed example, statistical methods are employed to determine the probability that a probable peak is an actual peak, not an actual peak, or that the data are too inconclusive to call.

Advantageously, the method and system for identifying a biological sample accurately makes composition calls in a highly automated manner. In such a manner, complete SNP profile information, for example, may be collected efficiently. More importantly, the collected data [is]are analyzed with highly accurate results. For example, when a particular composition is called, the result may be relied upon with great confidence. Such confidence is provided by the robust computational process employed[,] and the highly automatic method of collecting, processing, and analyzing the data set.

Please amend the paragraphs beginning on page 6, line 10, through page 7, line 20, as follows:

In accordance with the present invention, a method and device for identifying a biological sample is provided. Referring now to FIG. 1, an apparatus 10 for identifying a biological sample is disclosed. The apparatus 10 for identifying a biological sample generally comprises a mass spectrometer 15 communicating with a computing device 20. In a preferred embodiment, the mass spectrometer may be a MALDI-TOF mass spectrometer manufactured by

Bruker-Franzen Analytik GmbH; however, it will be appreciated that other mass spectrometers can be substituted. The computing device 20 is preferably a general purpose computing device. However, it will be appreciated that the computing device could be alternatively configured[,]; for example, it may be integrated with the mass spectrometer or could be part of a computer in a larger network system.

The apparatus 10 for identifying a biological sample may operate as an automated identification system having a robot 25 with a robotic arm 27 configured to deliver a sample plate 29 into a receiving area 31 of the mass spectrometer 15. In such a manner, the sample to be identified may be placed on the plate 29 and automatically received into the mass spectrometer 15. The biological sample is then processed in the mass spectrometer to generate data indicative of the mass of DNA fragments into biological sample. [This]These data may be sent directly to computing device 20, or may have some preprocessing or filtering performed within the mass spectrometer. In a preferred embodiment, the mass spectrometer 15 transmits unprocessed and unfiltered mass spectrometry data to the computing device 20. However, it will be appreciated that the analysis in the computing device may be adjusted to accommodate preprocessing or filtering performed within the mass spectrometer.

Referring now to FIG. 2, a general method 35 for identifying a biological sample is shown. In method 35, data [is]are received into a computing device from a test instrument in block 40. Preferably the data [is]are received in a raw, unprocessed and unfiltered form, but alternatively may have some form of filtering or processing applied. The test instrument of a preferred embodiment is a mass spectrometer as described above. However, it will be appreciated that other test instruments could be substituted for the mass spectrometer.

The data generated by the test instrument, and in particular the mass spectrometer, [includes]include information indicative of the identification of the biological sample. More specifically, the data [is]are indicative of the DNA composition of the biological sample. Typically, mass spectrometry data gathered from DNA samples obtained from DNA amplification techniques are noisier than, for example, those from typical protein samples. This is due in part because protein samples are more readily prepared in more abundance, and protein samples are more easily ionizable as compared to DNA samples. Accordingly, conventional mass spectrometer data analysis techniques are generally ineffective for DNA analysis of a biological sample.

Please amend the paragraphs beginning on page 8, line 13, through page 9, line 14 as follows:

Referring again to FIG. 2, the data received in block 40 [is]<u>are</u> denoised in block 45. The denoised data then has a baseline correction applied in block 50. A baseline correction is generally necessary as data coming from the test instrument, in particular a mass spectrometer instrument, has data arranged in a generally exponentially decaying manner. This generally exponential decaying arrangement is not due to the composition of the biological sample, but is a result of the physical properties and characteristics of the test instrument[,] and other chemicals involved in DNA sample preparation. Accordingly, baseline correction substantially corrects the data to remove a component of the data attributable to the test system[,] and sample preparation characteristics.

After denoising in block 45 and the baseline correction in block 50, a signal remains which is generally indicative of the composition of the biological sample. However, due to the extraordinary discrimination required for analyzing the DNA composition of the biological sample, the composition is not readily apparent from the denoised and corrected signal. For example, although the signal may include peak areas, it is not yet clear whether these "putative" peaks actually represent a DNA composition, or whether the putative peaks are result of a systemic or chemical aberration. Further, any call of the composition of the biological sample would have a probability of error which would be unacceptable for clinical or therapeutic purposes. In such critical situations, there needs to be a high degree of certainty that any call or identification of the sample is accurate. Therefore, additional data processing and interpretation [is]are necessary before the sample can be accurately and confidently identified.

Since the quantity of data resulting from each mass spectrometry test is typically thousands of data points, and an automated system may be set to perform hundreds or even thousands of tests per hour, the quantity of mass spectrometry data generated is enormous. To facilitate efficient transmission and storage of the mass spectrometry data, block 55 shows that the denoised and baseline corrected data [is]are compressed.

Please amend the paragraph on page 10, lines 1-11, as follows:

Referring again to block 40, data [is]are received from the test instrument, which is preferably a mass spectrometer. In a specific illustration, FIG. 3 shows an example of data from a mass spectrometer. The mass spectrometer data 70 generally comprises data points distributed along an x-axis [71] and a y-axis [72]. The x-axis [71] represents the mass of particles detected, while the y-axis [72] represents a numerical concentration of the particles. As can be seen in FIG. 3, the mass spectrometry data 70 is generally exponentially decaying with data at the left end of the x-axis [73] generally decaying in an exponential manner toward data at the heavier end [74] of the x-axis. However, the general exponential presentation of the data is not indicative of the composition of the biological sample, but is more reflective of systematic error and characteristics. Further, as described above and illustrated in FIG. 3, considerable noise exists in the mass spectrometry DNA data 70.

Please amend the paragraphs beginning on page 11, line 12, through page 12, line 3, as follows:

FIG. 5 shows an example of stage 0 high data 95. Since stage 0 high data 95 is generally indicative of the highest frequencies in the mass spectrometry data, stage 0 high data 95 will closely relate to the quantity of high frequency noise in the mass spectrometry data. In FIG. 6, an exponential fitting formula has been applied to the stage 0 high data 95 to generate a stage 0 noise profile 97. In particular, the exponential fitting formula is in the format $A_0 + A_1$ EXP $(-A_2m)$. It will be appreciated that other [expediential]exponential fitting formulas or other types of curve fits may be used.

Referring now to FIG. 7, noise profiles for the other high stages are determined. Since the later data points in each stage will likely be representative of the level of noise in each stage, only the later data points in each stage are used to generate a standard deviation figure that is representative of the noise content in that particular stage. More particularly, in generating the noise profile for each remaining stage, only the last five percent of the data points in each stage are analyzed to [determined]determine a standard deviation number. It will be appreciated that other numbers of points, or alternative methods could be used to generate such a standard deviation figure.

Please amend the paragraph on page 14, lines 10-12, as follows:

The formula [125] is generally indicated in FIG. 10. Once the signal has been denoised and shifted, a denoised and shifted signal 130 is generated as shown in FIG. 12. FIG. 11 shows an example of the wavelet coefficient 135 data set from the denoised and shifted signal 130.

Please amend the paragraph on page 15, lines 12-20, as follows:

Referring again to FIG. 2, the data from the baseline correction 50 is now compressed in block 55[,]; the compression technique used in a preferred embodiment is detailed in FIG. 18. In FIG. 18 the data in the baseline corrected data [is]are presented in an array format 182 with x-axis points 183 having an associated data value 184. The x-axis is indexed by the non-zero wavelet coefficients, and the associated value is the value of the wavelet coefficient. In the illustrated data example in table 182, the maximum value 184 is indicated to be 1000. Although a particularly advantageous compression technique for mass spectrometry data is shown, it will be appreciated that other compression techniques can be used. Although not preferred, the data may also be stored without compression.

Please amend the paragraphs on page 16, line 16, through page 17, line 14, as follows:

FIG. 19 generally describes the method of compressing mass spectrometry data, showing that the data file in block 201 is presented as an array of coefficients in block 202. The data starting point and maximum is determined as shown in block 203, and the intermediate real numbers are calculated in block 204 as described above. With the intermediate data points generated, the compressed data is generated in block 205. The described compression method is highly advantageous and efficient for compressing data sets such as a processed data set from a mass spectrometry instrument. The method is particularly useful for data, such as mass spectrometry data, that [uses]use large numbers and [has]have been processed to have occasional lengthy gaps in x-axis data. Accordingly, an x-y data array for processed mass spectrometry data may be stored with an effective compression rate of 10x or more. Although the compression technique is applied to mass spectrometry data, it will be appreciated with the method may also advantageously be applied to other data sets.

Referring again to FIG. 2, peak heights are now determined in block 60. The first step in determining peak height is illustrated in FIG. 20 where the signal 210 is shifted left or right to correspond with the position of expected peaks. As the set of possible compositions in the biological sample is known before the mass spectrometry data is generated, the possible positioning of expected peaks is already known. These possible peaks are referred to as expected peaks, such as expected peaks 212, 214, and 216. Due to calibration or other errors in the test instrument data, the entire signal may be shifted left or right from its actual position[,]; therefore, putative peaks located in the signal, such as putative peaks 218, 222, and 224 may be compared to the expected peaks 212, 214, and 216, respectively. The entire signal is then shifted such that the putative peaks align more closely with the expected peaks.

Please amend the paragraph on page 19, lines 3-14, as follows:

An indication of the confidence that each putative peak is an actual peak can be discerned by calculating a signal-to-noise ratio for each putative peak. Accordingly, putative peaks with a strong signal-to-noise ratio are generally more likely to be an actual peak than a putative peak with a lower signal-to-noise ratio. As described above and shown in FIG. 27, the height of each peak, such as height 272, 274, and 276, is determined for each peak, with the height being an indicator of signal strength for each peak. The noise profile, such as noise profile 97, is extrapolated into noise profile 280 across the identified peaks. At the center line of each of the peaks, a noise value is determined, such as noise value 282, 283, and 284. With a signal [values]value and a noise [values]value generated, signal-to-noise ratios can be calculated for each peak. For example, the signal-to-noise ratio for the first peak in FIG. 27 would be calculated as signal value 272 divided by noise value 282, and in a similar manner the signal-to-noise ratio of the middle peak in FIG. 27 would be determined as signal 274 divided by noise value 283.

Please amend the paragraph on page 21, lines 6-12, as follows:

In other situations, a more aggressive approach may be taken as sample data [is]are more pronounced or the risk of error may be reduced. In such a situation, the system may be set to assume a 100% probability with a 5 or greater target signal-to-noise ratio. Of course, an intermediate signal-to-noise ratio target figure can be selected, such as 7, when a moderate risk of error can be assumed. Once the target adjusted signal-to-noise ratio is set for the method, then for any adjusted signal-to-noise ratio a probability can be determined that a putative peak is an actual peak.

Please amend the paragraph on page 22, lines 6-15, as follows:

With the peak probability of each peak determined, the statistical probability for various composition components may be determined[. As], as an example, in order to determine the probability of each of three possible combinations of two peaks, -- peak G, peak C and combinations GG, CC and GC. FIG. 31 shows an example where a most probable peak 325 is determined to have a final peak probability of 90%. Peak 325 is positioned such that it represents a G component in the biological sample. Accordingly, it can be maintained that there is a 90% probability that G exists in the biological sample. Also in the example shown in FIG. 31, the second highest probability is peak 330 which has a peak probability of 20%. Peak 330 is at a position associated with a C composition. Accordingly, it can be maintained that there is a 20% probability that C exists in the biological sample.

Please amend the paragraphs on page 23, line 10, through page 24, line 18, as follows:

Once the probabilities of each of the possible combinations has been determined, FIG. 32 is used to decide whether or not sufficient confidence exists to call the genotype. FIG. 32 shows a call chart 335 which has an x-axis 337 which is the ratio of the highest combination probability to the second highest combination probability. The y-axis 339 simply indicates whether the ratio is sufficiently high to justify calling the genotype. The value of the ratio may be indicated by M [340]. The value of M is set depending upon trial data, sample composition, and the ability to accept error. For example, the value M may be set relatively high, such as to a value 4 so that the highest probability must be at least four times greater than the second highest probability before confidence is established to call a genotype. However, if a certain level of error may be acceptable, the value of M may be set to a more aggressive value, such as to 3, so that the ratio between the highest and second highest probabilities needs to be only a ratio of 3 or higher. Of course, moderate value may be selected for M when a moderate risk can be accepted. Using the example of FIG. 31, where the probability of GG was 72% and the probability of GC was 18%, the ratio between 72% and 18% is 4.0[,]; therefore, whether M is set to 3, 3.5, or 4, the system would call the genotype as GG. Although the preferred embodiment uses a ratio between the two highest peak probabilities to determine if a genotype confidently can be called, it will be appreciated that other methods may be substituted. It will also be appreciated that the above techniques may be used for calculating probabilities and choosing genotypes (or more general DNA patterns) consisting of combinations of more than two peaks.

Referring now to FIG. [32]33, a flow chart is shown generally defining the process of statistically calling genotype described above. In FIG. [32]33 block 402 shows that the height of each peak is determined and that in block 404 a noise profile is extrapolated for each peak. The signal is determined from the height of each peak in block [406]402 and the noise for each peak is determined using the noise profile in block [408]406. In block 410, the signal-to-noise ratio is calculated for each peak. To account for a non-Gaussian peak shape, a residual error is determined in block 412 and an adjusted signal-to-noise ratio is calculated in block 414. Block 416 shows that a probability profile is developed, with the probability of each peak existing found in block 418. An allelic penalty may be applied in block 420, with the allelic penalty applied to the adjusted peak probability in block 422. The probability of each combination of components is calculated in block 424 with the ratio between the two highest probabilities being determined in block 426. If the ratio of probabilities exceeds a threshold value, then the genotype is called in block 428.